# Modelling the asymmetric probabilistic delay of aircraft arrival 

J.V. Pérez-Rodríguez ${ }^{\text {a, * }}$, J.M. Pérez-Sánchez ${ }^{\text {b }}$, E. Gómez-Déniz ${ }^{\text {c }}$<br>${ }^{\text {a }}$ Department of Quantitative Methods for Economics \& Business, University of Las Palmas de Gran Canaria, Spain<br>${ }^{\mathrm{b}}$ Department of Quantitative Methods for Economics \& Business, University of Granada, Spain<br>${ }^{\text {c }}$ Department of Quantitative Methods in Economics and TiDES Institute, University of Las Palmas de Gran Canaria, Spain

## A R T I C L E I N F O

## Article history:

Received 31 October 2016
Received in revised form
6 March 2017
Accepted 7 March 2017

## Keywords:

Airports
Asymmetric link
Delay
Logit models
Bayesian estimation


#### Abstract

The main purpose of this paper is to present an asymmetric logit probability model to estimate and predict the daily probabilities of delay in aircraft arrivals. The proposed model takes into account statistical regularity, noting that more arrivals are on time than delayed, thus reflecting an asymmetric pattern of behaviour. The data analysed were obtained from the BTS and IATA databases for December 2014, corresponding to delays within the US airspace system for each carrier, measured at various US airports. The model was evaluated by analysing both in-sample and out-of-sample data, for main and control samples. The performance of the proposed asymmetric Bayesian logit model was compared with that of two others: frequentist logit and symmetric Bayesian logit. The main conclusion drawn is that the model we propose obtains the best fit, according to the statistics considered, and identifies a novel delaying factor, namely distance, which is not identified by the other models analysed.


© 2017 Elsevier Ltd. All rights reserved.

## 1. Introduction

Most studies addressing dichotomous outcomes, such as success vs. failure, use classical logit and probit models, and therefore assume that the responses are symmetric. Nevertheless, in practice, the real proportion of (for example) successes and failures may not be symmetric. If this is the case, application of these classical models can led to model misspecification and a misinterpretation of the marginal effects and unidentified predictors, the consequences of which could be very significant. In the present study, we examine data corresponding to aircraft arrival and departure delays, which often present just this kind of asymmetry.

Arrival and departure delays in the airspace system are important variables because they cause significant losses to airlines and create problems for passengers, airports and staff. Delays can be categorised into gate delay, taxi-out delay, en-route delay, terminal delay and taxi-in delay (see Mueller and Chatterji, 2002). Since traffic management decisions are influenced by the predicted demand, better demand forecasting is always desirable. Departure time uncertainty is the major cause of demand prediction error; therefore, increased departure time reliability will directly increase

[^0]the accuracy of demand prediction (Mueller and Chatterji, 2002). In consequence, scheduling and policy decision makers should seek to minimise the risk of delay and thus improve the forecasting accuracy of departure times when a probabilistic delay time model is used. Accordingly, it is important to determine the causes of delays in the airspace system, such as factors related to aircraft, airline operations, changes of procedure and traffic volume.

Several approaches can be taken to analyse this issue. On the one hand, we can attempt to estimate the actual duration of the delay. For example, Allan et al. (2001) analysed several determinants of flight delay at one US airport (Newark International Airport) and showed that adverse weather conditions influenced flight delays. On the other hand, Mueller and Chatterji (2002) modelled delay assuming it to be a random variable that follows a statistical distribution. Their study, seeking to improve delay prediction, analysed the departure, en-route and arrival delays of aircraft that operated out of one of ten major U.S. hub airports. Kwan and Hansen (2011) analysed causal factors including airport congestion, total traffic and en-route weather. The estimation results obtained suggested that airport congestion, measured by arrival queuing delay, was a major contributor to average delay (about 32\%). Nevertheless, these authors concluded that a model with a single explanatory variable is inadequate to describe the reality of a system. Wong and Tsai (2012) analysed flight delay propagation employing a survival method (the Cox proportional hazard model). These authors developed departure and arrival
delay models that showed how flight delay propagation can be formulated through repeated chain effects in aircraft rotations performed by a Taiwanese domestic airline. Other papers have also analysed delay propagation using other econometric methods; see, for example, Xu et al. (2005, 2007), Liu and Ma (2008) and Cao and Fang (2012), among others, who used a Bayesian network approach, Pyrgiotis et al. (2013), who analysed a network of airports using a queuing model, and Derudder et al. (2010) and Diana (2011), who analysed the prediction of arrival delays using spatial analysis.

Another possibility is to analyse the probability of delay. To our knowledge, only a few studies have taken this approach. Among them, Abdel-Aty et al. (2007) identified the periodic patterns of arrival delay for non-stop domestic flights at the Orlando International Airport during 2002-2003. Using logistic regression, their results showed that time of day, day of week, season, flight distance, precipitation at Orlando International Airport and scheduled time intervals between successive flights were significantly correlated with arrival delay. In this field, too, Tu et al. (2008) attempted to model flight departure delay probability distributions, but did not assess all of the relevant aviation and meteorological parameters. In another study, Wesonga et al. (2012) analysed the probability of arrival and departure delay at Entebbe airport (Uganda), using a multiple parametric approach to determine the probability of aircraft delay. In this study, a robust approach was used to include the apparently significant meteorological and aviation parameters while computing the exact probabilities of delay.

Motivated by the desire to improve the accuracy of demand prediction, both en-route and at airports, using probabilistic delay forecasting, we analyse departure and arrival data for U.S. airports with different volumes of traffic and significant delays. Following both Abdel-Aty et al. (2007) and Wesonga et al. (2012), we also use logistic regressions. Specifically, we analyse not only how the airport factor and the airline factor can influence delays, but also the distance between airports, departure delay and daily patterns. However, unlike the latter studies, our paper is conducted using an asymmetric logit model. This choice was made because it has been observed that the proportion of on-time flights (leaving/arriving within 15 min of the scheduled time) is generally higher than that of delayed ones.

This study examines data corresponding to air traffic delay statistics compiled in the United States, and uses an asymmetric logit model in the belief that this provides better results than the standard fitted logit model. ${ }^{1}$

The rest of this paper is organised as follows. Section 2 presents the methodology, including a brief description of the three probabilistic logit models analysed, with respect to aircraft delay: the classical logit model, the symmetric Bayesian logit model and the asymmetric Bayesian logit model. Section 3 presents the data included in the analysis and the sampling procedure. Section 4 presents the estimation performed, the main results obtained and their discussion. Finally, in Section 5 we summarise the main conclusions drawn.

[^1]
## 2. Methodology: logistic models

### 2.1. Classical models

Logistic regression has long been the standard method for studying the relationship between a binary response variable and one or more predictors or explanatory variables, using a cumulative density function (cdf), termed $\Psi$. Let $x$ be a vector of explanatory variables and $y$ the response variable taking values in $\{0,1\}$. This can be expressed as $\operatorname{Pr}(y=1 \mid x, \beta)=\Psi\left(x^{\prime} \beta\right)$. Then, by taking $\Psi$ as the cdf of the logistic distribution, we obtain the logistic regression. In this case, the probability density function is symmetric about zero. Thus, the cdf approaches 1 at the same rate as it approaches 0 . However, in many practical situations this is not a reasonable outcome, because data are often positively or negatively skewed and contain a substantial proportion of zeros (non-zeros) with respect to the proportion of non-zeros (zeros). This asymmetry may arise in diverse practical situations and then the logit model is not really appropriate. Since the pioneering study by Prentice (1976), various new models of dichotomous choice have been proposed to overcome this problem (under logit and probit assumptions), greatly assisted by advances in computer technology and software development. See, for example, Stukel (1988, 1990), Chen et al. (1999), Fernández and Steel (1998), Fletcher et al. (2005), Bazán et al. $(2006,2010)$ and Kumar and Manju (2015).

As observed by Stukel (1988) and Chen et al. (1999), the use of an asymmetric link function is recommended for binary response data when one response is much more frequent than the other.

The classical logit model is based on the following ideas. Let $y=$ $\left(y_{1}, y_{2}, \ldots, y_{n}\right)$ denote an $n \times 1$ vector of a dependent dichotomous variable, and let $x_{i}=\left(x_{i 1}, \ldots, x_{i k}\right)$ d denote the $k \times 1$ vector of covariates for the set $i$. Here, $x_{i 1}$ may be $\mathbf{1}$, which corresponds to an intercept. A fit regression model is used to estimate the probability of belonging to a group included in $y_{i}$. In this study of flight delays, if $y_{i}=1$ the $i$ th flight lands late, and $y_{i}=0$ otherwise. We assume that $y_{i}=1$ with probability $p_{i}$ and $y_{i}=0$ with probability $1-p_{i}$. The regression model is given by $p_{i}=F\left(x_{i}^{\prime} \beta\right)$, where $F$ is the inverse of the standard logistic cumulative function (link function), and $\beta=$ $\left(\beta_{1}, \ldots, \beta_{k}\right)$ is a $k \times 1$ vector of regression coefficients, which represents the effect of each variable $x_{i}$ on the model. Thus, the likelihood function, denoted as $l(y \mid x, \beta)$, is given by
$l(y \mid x, \beta)=\prod_{i=1}^{n}\left[F\left(x_{i}^{\prime} \beta\right)\right]^{y_{i}}\left[1-F\left(x_{i}^{\prime} \beta\right)\right]^{1-y_{i}}$,
where $F(s)=1 /\left(1+e^{-s}\right),-\infty<s<\infty$, is a symmetric function with respect to zero. Regression coefficients are usually estimated by numerical evaluation of the likelihood function. In the present case, thus, the model provides the probability of each flight landing with delay. The next step is to consider a cutoff in this probability in order to determine whether a flight will land on time or not. The logit model was evaluated using STATA econometric software.

### 2.2. Bayesian models allowing symmetry and asymmetry

The regression logit model outlined above is too simple to be used for any serious empirical work when the sample data present asymmetry between the two values of the binary response variable. In this context, the Bayesian approach is a powerful tool providing more flexible models in regression analysis.

The main idea of the Bayesian regression model (Zellner, 1971; Koop, 2003) is to consider that the regression coefficients are random and fit a distribution function (the prior distribution). We propose two alternative Bayesian estimations of the logit model.

Firstly, using a symmetric link function and secondly, an asymmetric link function, in which the first model appears as a special case.

From the asymmetric standpoint, an approach based on data augmentation (see Albert and Chib, 1993) can be used. In this case, it is easily shown that the asymmetric logit link is equivalent to considering the following:
$y_{i}= \begin{cases}1, & w_{i} \geq 0, \\ 0, & w_{i}<0,\end{cases}$
where $w_{i}=x_{i}^{\prime} \beta+\delta z_{i}+\varepsilon_{i}, z_{i} \sim G, \varepsilon_{i} \sim F$ and $i=1,2, \ldots, n$. We assume that $z_{i}$ and $\varepsilon_{i}$ are independent and that $F$ is the standard logistic cumulative distribution function. Moreover, $G$ is the cumulative distribution function of the half-standard normal distribution with pdf given by
$g\left(z_{i}\right)=\frac{2}{\sqrt{2 \pi}} e^{-z_{i}^{2} / 2}, \quad z_{i}>0$.
In this model, $\delta \in(-\infty, \infty)$ is the skewness parameter and so the skewness of the regression model is measured by $\delta z_{i}$. If $\delta>0$, the probability of $p_{i}=1$, i.e., the probability that the $i$ th flight will be delayed, increases. On the other hand, if $\delta<0$, the probability of it not being delayed increases. Obviously, the symmetric logit model is a special case of the previous model obtained for $\delta=0$.

The following likelihood function is thus obtained:
$l(y \mid x, \beta, \delta)=\prod_{i=1}^{n} \int_{0}^{\infty}\left[F\left(x_{i}^{\prime} \beta+\delta z_{i}\right)\right]^{y_{i}}\left[1-F\left(x_{i}^{\prime} \beta+\delta z_{i}\right)\right]^{1-y_{i}} g\left(z_{i}\right) d z_{i}$.

In the context of Bayesian analysis, a prior distribution must be specified for $\beta$ and $\delta$, say, $\pi(\beta, \delta)$. We assume non-informative and centred normal prior distributions for both parameters in order to facilitate comparison with frequentist estimations, i.e., $\beta_{j} \sim N\left(0, \sigma_{j}^{2}\right), \forall j=1, \ldots, k, \quad$ and $\quad \delta \sim N\left(0, \sigma_{\delta}^{2}\right), \quad$ considering $\sigma_{j}>0, \forall j=1, \ldots, k$, and $\sigma_{\delta}$ sufficiently large, noting the absence of prior knowledge about the parameters of interest, which facilitates comparison with the frequentist model.

By combining these prior assumptions with the likelihood shown in (4), we obtain the posterior distribution for the parameters $\beta$ and $\delta$, which is proportional to the prior times the likelihood,

Table 1
Descriptive statistics for variables used in the models.

| Variables (definition) | Mean | s.d. | Min. | Max. |
| :--- | :--- | :--- | :--- | :--- |
| Arrival delay (equals 1 if arrival delay) | 0.237 | 0.425 | 0 | 1 |
| Departure delay (equals 1 if departure delay) | 0.225 | 0.417 | 0 | 1 |
| Distance (between airports, in miles) | 815.18 | 602.35 | 31 | 4983 |
| LgAirLine (equals 1 if large airline) | 0.535 | 0.499 | 0 | 1 |
| MedAirLine (equals 1 if medium-sized airline) | 0.333 | 0.471 | 0 | 1 |
| SmAirLine (equals 1 if small airline) | 0.131 | 0.338 | 0 | 1 |
| LgAirport (equals 1 if large airport) | 0.236 | 0.425 | 0 | 1 |
| MedAirport (equals 1 if medium-sized airport) | 0.181 | 0.385 | 0 | 1 |
| SmAirport (equals 1 if small airport) | 0.106 | 0.308 | 0 | 1 |
| RestAirport (equals 1 if other airports) | 0.476 | 0.499 | 0 | 1 |
| Tuesday (equals 1 if Tuesday) | 0.163 | 0.369 | 0 | 1 |
| Wednesday (equals 1 if Wednesday) | 0.153 | 0.360 | 0 | 1 |
| Thursday (equals 1 if Thursday) | 0.130 | 0.337 | 0 | 1 |
| Friday (equals 1 if Friday) | 0.138 | 0.345 | 0 | 1 |
| Saturday (equals 1 if Saturday) | 0.115 | 0.319 | 0 | 1 |
| Sunday (equals 1 if Sunday) | 0.129 | 0.335 | 0 | 1 |
| Monday (equals 1 if Monday) | 0.170 | 0.376 | 0 | 1 |

component wise Metropolis Hasting algorithm (a Markov-Chain simulation method) that provides a posterior simulation for $\pi(\beta, \delta \mid y, x)$. We can sample ( $\beta, \delta$ ) from this posterior distribution using the WinBUGS package, ${ }^{2}$ based on Gibbs sampling applying Markov Chain Monte Carlo (MCMC) methods (see Carlin and Polson, 1992; Gilks et al., 1995, for further details).

## 3. Data

### 3.1. Databases

In this study, we consider two sets of data. First, the airline on-time performance data obtained from the Bureau of Transportation Statistics (BTS), which compiles delay data for the benefit of passengers in the USA. ${ }^{3}$

In particular, we focused on information related to arrivals at U.S. airports in December 2014 ( 477183 observations). This database contains characteristics associated with origin and destination airports, airline names, arrival and departure delays, distance (in miles) between airports, the day of the week (we considered six dummies and the reference category, Mondays), among other factors. A delayed flight is defined as one in which the aircraft fails to release its parking brake within 15 min of the scheduled departure time. ${ }^{4}$

The second database comprises passenger traffic information
$\pi(\beta, \delta \mid y, x) \propto l(y \mid x, \beta, \delta) \pi(\beta, \delta)=\left\{\prod_{i=1}^{n} \int_{0}^{\infty}\left[F\left(x_{i}^{\prime} \beta+\delta z_{i}\right)\right]^{y_{i}}\left[1-F\left(x_{i}^{\prime} \beta+\delta z_{i}\right)\right]^{1-y_{i}} g\left(z_{i}\right) d z_{i}\right\} \pi(\beta, \delta)$.

This posterior distribution summarises all the information, both prior and data-based, possessed about the unknown parameters, $\beta$ and $\delta$.

In simple Bayesian models, it is usually easy to derive the posterior distribution directly when conjugate prior distributions are used. However, for more complex assumptions, we need to factor the posterior distribution and simulate it in parts, generally the marginal posterior distribution of the parameters (or hyperparameters), and then simulate the other parameters conditional on the data and the simulated parameters. This procedure is facilitated by using the Gibbs sampler algorithm, which is a special
published by IATA (the International Air Transport Association) with data from 2013 including company size and airport size. From

[^2]

Fig. 1. Density functions for aircraft arrivals/departures.


Fig. 2. Daily proportions of delay and on-time.
these data, we constructed the variables as follows. First, the airlines were divided into three groups, according to the size data published for 2013: the variable "large airlines" was assigned the value one for companies with 60,000-120,000 scheduled passengers (Delta Air Lines, Southwest Airlines, United Airlines, American Airlines, China Southern Airlines, Ryanair, China Eastern Airlines and Lufthansa). The variable "medium-sized airlines" was assigned the value one for airlines with $20,000-60,000$ passengers (31 companies). The reference category for "small airlines", assigned the value 0 , included the remaining companies. Airport size was also expressed in four categories, using the 2013 data: "large airports" were defined as those with over 60 million scheduled passengers (Atlanta, Chicago-O’Hare, Los Angeles, Dallas-Fort Worth and Denver); "medium-sized airports" received 40 to 60 million passengers (New York JFK, San Francisco, Las Vegas, Phoenix-Sky Harbor, Houston-George Bush, Charlotte-Douglas, Miami and Orlando-International), and "small airports" had 25 to 40 million passengers (Newark, Minneapolis-Saint Paul, Seattle-Tacoma, Detroit-Wayne, Philadelphia and Boston). The reference category, in this case, included all remaining airports (those with fewer than 25 million passengers).

Table 1 shows the descriptive statistics obtained for the variables included in each of the models considered. Many of them are dichotomous variables and one is continuous (distance, in miles, between airports). For the dichotomous variables, the mean represents the proportion of ones in the variable. For example, 23.7\% and $22.5 \%$ represent the proportion of arrival and departure delays on December 2014, respectively. The average distance between airports was 815.18 miles.

Fig. 1 shows that density functions for departure and arrival
delay are asymmetric to the right and that the mode is equal to zero minutes; in other words, that the aircraft predominantly arrive on time. The pattern observed for aircraft arriving at and departing from other airports was similar to that shown in Fig. 2 (interday delay and on-time patterns for arrivals and departures). The average proportions of delay were around 0.22 on departure and 0.23 on arrival, i.e., 0.78 and 0.77 are the on-time proportions for departures and arrivals, respectively. The main conclusion we draw from these findings is that the interday patterns for arrival and departure delays are not uniformly distributed ${ }^{5}$ and that arrivals and departures are more likely to be delayed on a Tuesday than on any other day in the month analysed. ${ }^{6}$

### 3.2. Sampling procedure

The two databases were matched and the models were estimated using the following sampling procedure.

Due to the computational burden of dealing with 477183

[^3]observations for one month using Markov Chain Monte Carlo (MCMC), we decided to randomly select two samples for comparison purposes. First, we named as the main sample one formed by 90,000 randomly chosen observations. Second, we constructed another randomly selected sample formed of the same number of observations. This sample was named the control sample.

In both the main and the control samples, three models were estimated: standard (or frequentist), symmetric non-informative Bayesian and asymmetric non-informative Bayesian models. The estimation in these samples was considered in-sample.

To evaluate the predictive ability of the models estimated, we randomly selected another sample, termed the out-of-sample, which was formed by 5000 randomly selected observations, none of which formed part of the in-sample observations.

The following procedure was then applied: 1) Estimate the parameters of the in-sample models. 2) Use the estimated parameters in the out-of-sample to construct a measure of predictive ability.

### 3.3. Some causal factors on the probability of arrival delay

A flight is considered delayed when it arrived 15 or more minutes after the scheduled time. Delayed minutes are calculated for delayed flights only. Several variables were employed as flight probability delay factors. Previous studies have shown that arrival delays may be related to departure delays, the time period (weekday or weekend) and total traffic and passengers.

Following Abdel-Aty et al. (2007) we considered two variables. First, the day of the week was used to investigate the daily pattern (deterministic). Second, we considered the distance (in miles) between airports, in order to investigate the effect of long and short flight distances on arrival delay.

Other variables used were related to passenger numbers at airports and transported by airlines, and also whether there were arrival delays associated with departure delays. The latter can be considered a proxy variable of the propagation delay. For example, arrival delays could be highly correlated with regard to earlier flights (departure delays). Given that the punctuality of one flight is sensitive to earlier flight delays - because delays tend to propagate over time - the departure delay must be taken into account as a factor related to the propagation delay. Therefore, departure delay was taken as a dummy variable with a value of 1 if the departure delay was 15 min later than the scheduled time. Following BTS, such a delay might be caused by meteorological factors (i.e., significant meteorological conditions, actual or forecast, that delayed or prevented the operation of a flight, such as tornado, blizzard or hurricane); those attributable to the national aviation system (i.e., non-extreme weather conditions, airport operations, heavy traffic volume and air traffic control factors); other operational factors that are the airline's responsibility (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fuelling, etc.); or delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of a security breach, inoperative screening equipment and/or long queues (in excess of 29 min ) at screening areas. Some of these variables are highly correlated with departure delay. For example, the weather conditions variable has a linear correlation coefficient equal to 0.82 with departure delay. Accordingly, the weather conditions variable was omitted because this information is implicity included in the departure delay variable.

Finally, we included two variables representing the volume of passengers. First, the airport type, regarding the number of passengers, which is a categorical variable that allows us distinguish
the size of the airport in terms of the volume of passengers at each airport. Second, the airline type, by number of passengers, which is another categorical variable and allows us to indicate the importance of travelling with airlines carrying large (or small) numbers of passengers.

## 4. Results

### 4.1. Estimation and prediction results

The models for both the main and the control samples were estimated. The results obtained in both samples were robust, with no significant differences, in terms of signs or relevant factors, and therefore we focused on the main sample. For this reason, the results obtained for the control sample are shown in the Appendix (Table 5).

Table 2 shows the in-sample estimates for the main sample, obtained by each of the three models: standard (or frequentist), symmetric non-informative Bayesian and asymmetric non--informative Bayesian.

In each of these estimates, non-informative prior distributions were assumed, i.e., $\sigma_{j}^{2}=10^{8}, \forall j=1, \ldots, k$, and $\sigma_{\delta}^{2}=10^{8}$. The posterior distributions for Bayesian models were simulated using WinBUGS in two samples. A total of 500,000 iterations were carried out (after a burn-in period of 100,000 simulations) for each sample. Three different chains were performed and the convergence was evaluated for all parameters using tests provided within the WinBUGS Convergence Diagnostics and Output Analysis software.

To assess the goodness of fit and the forecast obtained by the frequentist and the Bayesian logit models, four different measures were applied: (1) the percentage of correct fits; (2) the percentage of correct predictions; (3) three statistical fit measures, the deviance (DIC), the Akaike (AIC) and the Bayesian (BIC) information criterions; and (4) the $c$-statistic (area under the ROC curve) which measures the goodness of fit in the logistic model curve.

As expected, because the prior information is non-informative, the symmetric Bayesian results were very similar to those obtained by the standard frequentist estimations. As shown in Table 2, according to the frequentist and symmetric Bayesian estimations, the intercept, departure delays, large airlines, medium-sized airlines, large airports, medium-sized airports, Tuesday flights, Saturday flights and Sunday flights are all statistically significant in explaining the probability of arrival delay of a flight, at the $1 \%$ significance level. Only the Friday flights appear (at 10\%) as a new positive factor in the symmetric Bayesian estimations. With respect to the positive relationship, the greater the departure delay, the greater the probability of an arrival delay; in addition, large and medium-sized airports are more prone to delays than small ones (those with fewer than 25 million passengers) and there is more probability of delay on a Tuesday than a Monday during the analysed month. Moreover, large and medium-sized airlines have a lower probability of delay than small ones, and this probability also decreases at weekends (with respect to Mondays). It is noteworthy that the distance between airports is not a significant factor in delay probability, according to the frequentist and symmetric Bayesian models.

In the asymmetric Bayesian estimations, the $\delta$ parameter, which indicates the possible asymmetry of the model, is statistically significant. This model detected the same results as the frequentist and symmetric Bayesian ones in terms of relevant factors, but also detected two new ones: the greater the distance between airports, the greater the probability of a delay occurring (at $1 \%$ significance)

Table 2
In-sample logit estimation results for each model in the main sample.

| Variables | Frequentist |  |  | Symmetric Bayesian |  |  | Asymmetric Bayesian |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\widehat{\beta}$ | Robust s.e. | $p$-value | $\widehat{\beta}$ | MC error | s.d. | $\widehat{\beta}$ | MC error | s.d. |
| Intercept | $-2.280^{* * *}$ | 0.044 | 0.000 | $-2.280^{* * *}$ | 0.000 | 0.044 | $4.699^{* * * *}$ | 0.036 | 0.457 |
| Departure delay | 4.121*** | 0.024 | 0.000 | $4.123^{* * *}$ | 0.000 | 0.024 | $55.38{ }^{* * *}$ | 0.377 | 4.504 |
| Distance | 5. $10^{-6}$ | $2 \cdot 10^{-5}$ | 0.800 | $5 \cdot 10^{-6}$ | 0.000 | $2 \cdot 10^{-5}$ | 9.10 ${ }^{-4 * * *}$ | 7-10-6 | $10^{-4}$ |
| LgAirLine | $-0.664^{* *}$ | 0.035 | 0.000 | $-0.664^{* * *}$ | 0.000 | 0.035 | $-3.162^{* * *}$ | 0.022 | 0.301 |
| MedAirLine | $-0.242^{* *}$ | 0.036 | 0.000 | $-0.243^{* * *}$ | 0.000 | 0.036 | $-1.311^{* * *}$ | 0.014 | 0.230 |
| LgAirport | $0.108 * * *$ | 0.029 | 0.000 | $0.108 * * *$ | 0.000 | 0.029 | $0.813^{* * *}$ | 0.006 | 0.148 |
| MedAirport | $0.234^{* * *}$ | 0.032 | 0.000 | 0.234*** | 0.000 | 0.032 | 1.287*** | 0.008 | 0.182 |
| SmAirport | -0.001 | 0.041 | 0.967 | -0.001 | 0.000 | 0.041 | 0.286* | 0.004 | 0.182 |
| Tuesday | 0.297*** | 0.039 | 0.000 | $0.297^{* * *}$ | 0.000 | 0.039 | $1.358^{* * *}$ | 0.011 | 0.221 |
| Wednesday | 0.028 | 0.041 | 0.492 | 0.028 | 0.000 | 0.041 | 0.104 | 0.007 | 0.190 |
| Thursday | -0.048 | 0.043 | 0.262 | -0.048 | 0.000 | 0.043 | -0.222 | 0.007 | 0.192 |
| Friday | 0.063 | 0.041 | 0.121 | 0.063* | 0.000 | 0.041 | 0.133 | 0.007 | 0.196 |
| Saturday | $-0.167^{* * *}$ | 0.045 | 0.000 | -0.168*** | 0.000 | 0.045 | $-0.722^{* * *}$ | 0.007 | 0.200 |
| Sunday | $-0.168^{* * *}$ | 0.044 | 0.000 | $-0.168^{* * *}$ | 0.000 | 0.043 | $-0.570^{* * *}$ | 0.007 | 0.196 |
| $\delta$ |  |  |  |  |  |  | $-44.99^{* * *}$ | 0.307 | 3.676 |
| DIC | 54195.07 |  |  | 54223.10 |  |  | 15351.90 |  |  |
| AIC | 54223.07 |  |  | 54240.0 |  |  | 8658.0 |  |  |
| BIC | 54354.78 |  |  | 54370.0 |  |  | 8799.0 |  |  |
| \% correct fit | 90.36 |  |  | 90.36 |  |  | 100 |  |  |
| \% correct prediction | 90.56 |  |  | 90.56 |  |  | 90.56 |  |  |

*** indicates $1 \%$ significance or relevance level.
${ }^{* *}$ indicates $5 \%$ significance or relevance level.

* indicates $10 \%$ significance or relevance level.
and the fact of landing at a small airport increases the probability of delay (at $10 \%$ significance). As expected, the signs of the relevant factors are unchanged but the estimated coefficients differ considerably from those of the other two models. This difference is further accentuated in the estimation of the intercept. Furthermore, the parameter $\delta$, which measures the skewness of the data, is statistically significant and negative at the $1 \%$ level of significance, after adjusting the estimated probability of delay, i.e, decreasing this probability and correcting the evident asymmetry in the data. We believe that in the first two models the estimated intercept may contain part of the asymmetry effect made apparent in the asymmetric model. The results for the control sample (Table 5) are similar to these, in sign and relevant factors, but the Friday flight becomes positive again for delay.

For our database, we obtained a DIC of 54223.07 for the frequentist logit model and a correct fit rate of $90.36 \%$. With Bayesian estimation, the deviance information criterion (DIC) measures were 54223.10 and 15351.90 for the symmetric and asymmetric models, respectively, i.e, the DIC value obtained for the asymmetric model was notably lower than the one obtained for the non-asymmetric models. The same pattern was obtained for the other criteria. The major reduction in these measures indicates a significant increase in the level of fit. Furthermore, the asymmetric model obtained better classification results, with $100 \%$ correct fit. Thus, the leverage of this model is much better than that of the

Table 3
In-sample descriptive statistics for the estimated probabilities in the main sample.

|  | Frequentist | Symmetric Bayesian | Asymmetric Bayesian |
| :--- | :--- | :--- | :--- |
| mean | 0.2402 | 0.2368 | 0.2368 |
| s.d. | 0.3124 | 0.3112 | 0.3946 |
| skewness | 1.3201 | 1.3214 | 1.2854 |
| kurtosis | 2.7912 | 2.7909 | 2.7093 |
| $p_{25}$ | 0.0593 | 0.0538 | 0.0131 |
| $p_{50}$ | 0.0789 | 0.0762 | 0.0212 |
| $p_{75}$ | 0.1207 | 0.1201 | 0.0479 |
| $p_{90}$ | 0.8256 | 0.8153 | 0.9916 |
| $c-$ statistic | 0.8793 | 0.8806 | 1.0000 |

symmetric models. Finally, the measure of correct predictions shows that all three models performed well, with an overall value of $90.56 \%$. The threshold probability used to fit and predict a delay was the sample frequency of delay, namely 0.237 ( 0.234 for the control sample).

Table 3 shows the mean, standard deviation, skewness, kurtosis and some percentiles for the estimated probabilities obtained by the models. This table also shows the $c$-statistics (area under the ROC curve) obtained. The best fit is again obtained with the asymmetric Bayesian model.

### 4.2. Discussion of results

In this study, several logistic regression models were estimated by employing data from the US airspace system for each carrier at different US airports, combined with IATA data. Three models were analysed: first, the frequentist and the non-informative symmetric Bayesian models. When the prior distribution is non-informative, these two models conclude in the same way. Both are based exclusively on the sample information and the estimations obtained (for the main and control samples), logically, are very similar. The third model, proposed in this paper, is the non-informative asymmetric Bayesian model. The results obtained indicate that this latter model produces a better in-sample fit than the other symmetric models, while the predicting capacity remains unchanged as regards the out-of-sample fit.

Table 4
Airports and airlines' estimated probabilities under the asymmetric Bayesian model in the main sample.

| Airport/Airline | $N$ | $\overline{\operatorname{Pr}}(y=1)$ | s.d. |
| :--- | :--- | :--- | :--- |
| Large airports | 21251 | 0.224 | 0.384 |
| Medium-sized airports | 16339 | 0.260 | 0.408 |
| Small airports | 9582 | 0.214 | 0.377 |
| Other airports | 42828 | 0.239 | 0.397 |
| Large airlines | 48140 | 0.230 | 0.391 |
| Medium-sized airlines | 30017 | 0.231 | 0.389 |
| Small airlines | 11843 | 0.277 | 0.417 |



Fig. 3. Average estimated probability of delay, by airport and day of week.

Several factors, proposed in earlier studies, were considered relevant to the probability of delay (see subsection 3.3). Departure delay was the most significant factor obtained in our models, and
therefore we conclude that prior flight delays present a high degree of dependence (propagation delay). In prior research, this variable is generally omitted from probability models. However, as it is


Fig. 4. Average estimated probability of delay, by airline and day of week.
correlated with meteorological factors (see subsection 3.3), this variable is a significant factor in explaining the probability of arrival delay. Second, regarding the coefficient of distance, our results indicate that this factor has a positive impact on probability of arrival delay with odd-ratio of $\exp \left(9.10^{-4}\right)=1.001$ which indicates that by increasing one mile, the probability changes very little. Abdel-Aty et al. (2007) considered the distance between regressors to estimate the pattern of single flight delay as a categorical variable, using the following distances: less than 750 miles, 750-1000 miles and greater than 1000 miles. These authors found that the odds of delay for a flight with a distance of 750-1000 miles were higher than those for flights with other ranges. This may be because longer flights have more opportunities to save time during the flight and thus avoid arrival delay. Third, regarding the coefficients of airport type and airline type, we found that large airlines had a 96 times greater probability of on-time arrival than small ones $(\exp (-3.162)=0.04)$ and that medium-sized airports had a 262.2 times greater probability of arrival delay than those with fewer than 25 million passengers $(\exp (1.287)=3.622)$. Table 4 shows the estimated probabilities $(\overline{\operatorname{Pr}}(y=1))$ under asymmetric Bayesian estimation, taking into account the sizes of the airports and airlines. $N$ represents the number of observations and s.d. is the standard deviation of the estimated probabilities. The average probability of arrival delay is higher for medium-sized airports and small airlines. However, small airports and large airlines have lower average probabilities. Finally, regarding the day of the week coefficients, results indicate that Tuesday is the most significant variable. The odds ratio for this day is $\exp (1.287)=3.62$ which is higher than that for the other time periods. This finding supports a daily pattern of delay, which was also reported by AbdelAty et al. (2007). However, these authors found Friday to be the most significant in their sample, possibly because December 24th (Christmas Eve) and 31st were on a Wednesday and the demand was more intensive than on the Tuesday. Figs. 3 and 4 show the average asymmetric Bayesian estimated probabilities of delay by day of the week, for airports and airlines, respectively. In both cases, the probability is higher on Tuesdays, with the exception of small airlines (Friday).

## 5. Summary and conclusions

This study analyses the use of an asymmetric Bayesian logit model to estimate the probability of aircraft delay, taking into account the asymmetric pattern of arrival delays at U.S. airports. To the best of our knowledge, asymmetric Bayesian logit models have not previously been applied in this setting and with these intentions.

We evaluated this model by comparing its results with those obtained by the frequentist and symmetric Bayesian approaches.

The main results obtained show that, according to the frequentist and standard Bayesian logit methods, the departure delay, the size of the airline, the size of the airport and the day of the flight (Tuesdays and weekends) are statistically significant factors (at the $1 \%$ significance level) to explain the probability of delay. Our study shows that arrival delay is strongly related to the origin-departure delay. The latter delay is attributed to operating procedures (i.e., the first flight segment of the day typically departs late).

In our asymmetrical Bayesian model, we also identify an important new delay factor with respect to the frequentist and symmetric Bayesian models, namely the distance, in miles, between airports (statistically significant at $1 \%$ ). Furthermore, the importance of incorporating asymmetry into the model is clearly corroborated by the information criteria, the percentage of correct fit and the $c$-statistic based on the ROC curve.

## Acknowledgements

José María Pérez-Sánchez and Emilio Gómez-Déniz were partially supported by the project ECO2013-4702 (MINECO, Spain).

The authors thank the anonymous reviewers for the detailed and helpful comments made regarding this manuscript.

## Appendix

Table 5
In-sample logit estimation results for each model in the control sample

| Variables | Frequentist |  |  | Symmetric Bayesian |  |  | Asymmetric Bayesian |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\widehat{\beta}$ | Robust s.e. | $p$-value | $\widehat{\beta}$ | MC error | s.d. | $\widehat{\beta}$ | MC error | s.d. |
| Intercept | $-2.280^{* * *}$ | 0.044 | 0.000 | $-2.284^{* * *}$ | 0.000 | 0.043 | -0.674 | 0.160 | 2.856 |
| Departure delay | $4.123^{* * *}$ | 0.024 | 0.000 | $4.125^{* *}$ | 0.000 | 0.024 | $16.05^{* * *}$ | 0.775 | 13.78 |
| Distance | $2 \cdot 10^{-5}$ | $2 \cdot 10^{-5}$ | 0.398 | $10^{-5}$ | 0.000 | $1 \cdot 10^{-5}$ | 3.10 ${ }^{-4 * * *}$ | $2 \cdot 10^{-5}$ | 3. $10^{-5}$ |
| LgAirLine | $-0.706^{* * *}$ | 0.035 | 0.000 | $-0.700^{* *}$ | 0.000 | 0.035 | $-1.329^{* *}$ | 0.038 | 0.685 |
| MedAirLine | $-0.286^{* * *}$ | 0.036 | 0.000 | $-0.285^{* *}$ | 0.000 | 0.036 | $-0.575^{* * *}$ | 0.018 | 0.336 |
| LgAirport | $0.086^{* *}$ | 0.030 | 0.004 | $0.086^{* * *}$ | 0.000 | 0.029 | $0.291^{* * *}$ | 0.012 | 0.230 |
| MedAirport | $0.233^{* * *}$ | 0.031 | 0.000 | $0.233^{* * *}$ | 0.000 | 0.032 | $0.435^{* * *}$ | 0.012 | 0.228 |
| SmAirport | 0.020 | 0.041 | 0.613 | 0.020 | 0.000 | 0.040 | 0.123 | 0.006 | 0.141 |
| Tuesday | $0.265^{* * *}$ | 0.039 | 0.000 | $0.265^{* * *}$ | 0.000 | 0.039 | $0.506^{* * *}$ | 0.014 | 0.273 |
| Wednesday | 0.041 | 0.040 | 0.308 | 0.041 | 0.000 | 0.040 | 0.090 | 0.003 | 0.100 |
| Thursday | -0.011 | 0.043 | 0.796 | -0.011 | 0.000 | 0.043 | -0.015 | 0.001 | 0.091 |
| Friday | $0.097^{* *}$ | 0.041 | 0.018 | 0.097*** | 0.000 | 0.041 | $0.136^{* *}$ | 0.002 | 0.099 |
| Saturday | $-0.176^{* * *}$ | 0.044 | 0.000 | $-0.176^{* * *}$ | 0.000 | 0.045 | $-0.363^{* *}$ | 0.011 | 0.219 |
| Sunday | $-0.116^{* * *}$ | 0.043 | 0.007 | $-0.117^{* * *}$ | 0.000 | 0.043 |  | $0.004$ | $0.117$ |
| $\delta$ |  |  |  |  |  |  | $-10.58^{* * *}$ | 0.747 | 1.329 |
| DIC | 53909.76 |  |  | 53937.7 |  |  | -29377.30 |  |  |
| AIC | 53937.76 |  |  | 53950.0 |  |  | 33134.00 |  |  |
| BIC | 54069.46 |  |  | 54080.0 |  |  | 33275.11 |  |  |
| \% correct fit | 90.38 |  |  | 90.38 |  |  | 100 |  |  |
| \% correct prediction | 90.56 |  |  | 90.56 |  |  | 90.56 |  |  |

[^4]
## References

Abdel-Aty, M., Lee, C., Bai, Y., Li, X., Michalak, M., 2007. Detecting periodic patterns of arrival delay. J. Air Transp. Manag. 13 (6), 355-361.
Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. 88 (422), 669-679.
Allan, S.S., Beesley, J.A., Evans, J.E., Gaddy, S.G., 2001. Analysis of delay at newark international airport. In: Proceedings of 4th USA/Europe Air Traffic Management R\&D Symposium. Santa Fe, NM. 3-7 December, 2001.
Bazán, J.L., Branco, M.D., Bolfarine, H., 2006. A skew item response model. Bayesian Anal. 1 (4), 861-892.
Bazán, J.L., Bolfarine, H., Branco, M.D., 2010. A framework for skewed-probit links in binary regression. Commun. Statistics-Theory Methods 39, 678-697.
Bermúdez, L.L., Pérez-Sánchez, J.M., Ayuso, M., Gómez-Déniz, E., Vázquez, F.J., 2008. A Bayesian dichotomous model with asymmetric link for fraud in insurance. Insurance: Math. Econ. 42, 779-786.
Cao, W., Fang, X., 2012. Airport flight departure delay model on improved BN structure learning. Phys. Procedia 33, 597-603.
Carlin, B.P., Polson, N.G., 1992. Monte Carlo Bayesian methods for discrete regression models and categorical time series. Bayesian Stat. 4, 577-586.
Chen, M.H., Dey, D.K., Shao, Q.M., 1999. A new skewed link model for dichotomous quantal response data. J. Am. Stat. Assoc. 94, 1172-1186.
Derudder, B., Devriendt, L., Witlox, F., 2010. A spatial analysis of multiple airport cities. J. Transp. Geogr. 18, 345-353.
Diana, T., 2011. Predicting arrival delays: an application of spatial analysis. J. Aircr. 48 (2), 462-467.
Fernández, C., Steel, M.F.J., 1998. On Bayesian modelling of fat tails and skewness. J. Am. Stat. Assoc. 93, 359-371.

Fletcher, D., Mackenzie, D., Villouta, E., 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. Environ. Ecol. Stat. 12, 45-54.
Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1995. Introducing Markov chain Monte Carlo. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman and Hall, London.
Koop, G., 2003. Bayesian Econometrics. Wiley, Chichester.
Kumar, C.S., Manju, L., 2015. On modified skew logistic regression model and its applications. Statistica 75 (4), 361-377.
Kwan, I., Hansen, M., 2011. US flight delay in the 2000s: an econometric analysis. In:

Transportation Research Board 90th Annual Meeting. No. 11-4283.
Liu, Y.J., Ma, S., 2008. Flight delay and delay propagation analysis based on Bayesian network. In: International Symposium on Knowledge Acquisition and Modeling. KAM'08, pp. 318-322 (IEEE).
Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS: a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 10, 325-337.
Mueller, E.R., Chatterji, G.B., 2002. Analysis of Aircraft Arrival and Departure Delay Characteristics. AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical. NASA Ames Research Center, Moffett Field, Los Angeles.
Pérez-Sánchez, J.M., Negrín-Hernández, M.A., García-García, C., Gómez-Déniz, E., 2014. Bayesian asymmetric logit model for detecting risk factors in motor ratemaking. ASTIN Bull. 44 (2), 445-457.
Prentice, R.L., 1976. Generalization of the probit and logit methods for dose response curves. Biometrics 32, 761-768.
Pyrgiotis, N., Malone, K.M., Odoni, A., 2013. Modelling delay propagation within an airport network. Transp. Res. Part C Emerg. Technol. 27, 60-75.
Sáez-Castillo, A.J., Olmo-Jiménez, M.J., Pérez, J.M., Negrín, M., Arcos, A., Díaz, J., 2010. Bayesian analysis of nosocomial infection risk and length of stay in a department of general and digestive surgery. Value Health 13 (4), 431-439.
Stukel, T., 1988. Generalized logistic model. J. Am. Stat. Assoc. 83, 426-431.
Stukel, T., 1990. A general model for estimating $\mathrm{ED}_{100}$ p for binary response dos-e-response data. Am. Stat. 44 (1), 19-22.
Tu, Y., Ball, M., Jank, W., 2008. Estimating flight departure delay distributions v A statistical approach with long-term trend and short-term pattern. J. Am. Stat. Assoc. 103, 112-125.
Wesonga, R., Nabugoomu, F., Jehopio, P., 2012. Parameterized framework for the analysis of probabilities of aircraft delay at an airport. J. Air Transp. Manag. 23, 1-4.
Wong, J.-T., Tsai, S.-C., 2012. A survival model for flight delay propagation. J. Air Transp. Manag. 23, 5-12.
Xu, N., Laskey, K.B., Donohue, G., Chen, C.H., 2005. Estimation of delay propagation in the national aviation system using Bayesian networks. In: 6th USA/Europe Air Traffic Management Research and Development Seminar.
Xu, N., Laskey, K.B., Chen, C.C., Williams, S.C., Sherry, L., 2007. Bayesian network analysis of flight delays. In: TRB 2007 Annual Meeting. CD-ROM.
Zellner, A., 1971. An Introduction to Bayesian Inference in Econometrics. Wiley, New York.


[^0]:    * Corresponding author.

    E-mail addresses: jv.perez-rodriguez@ulpgc.es (J.V. Pérez-Rodríguez), josemag@ugr.es (J.M. Pérez-Sánchez), emilio.gomez-deniz@ulpgc.es (E. Gómez-Déniz).

[^1]:    ${ }^{1}$ This approach has been successfully used in other studies; for example, Chen et al. (1999) applied a Bayesian approach and an asymmetric link in analysing binary response data, when one response is much more frequent than the other. Similarly, Bermúdez et al. (2008) applied asymmetric logistic regression to model fraudulent behaviour, using a Spanish insurance database. In the area of health care, Sáez-Castillo et al. (2010) used an asymmetric logistic link to predict infection rates in a General and Digestive Surgery hospital department. More recently, Pérez-Sánchez et al. (2014) analysed the risk factors of automobile insurance claims, considering an asymmetric link in the logistic regression.

[^2]:    ${ }^{2}$ Windows Bayesian inference using Gibbs Sampling, developed jointly by the MRC Biostatistics Unit [University of Cambridge, Cambridge, UK] and the Imperial College School of Medicine at St. Mary's, London (Lunn et al., 2000).
    ${ }^{3}$ The data used here can be obtained from: http://www.transtats.bts.gov/, the Office of the Assistant Secretary for Research and Technology (OST-R).
    ${ }^{4}$ The FAA is more interested in delays indicating surface movement inefficiencies and will record a delay when an aircraft requires 15 min or longer over the standard taxi-out or taxi-in time (Out to Off time, or On to In time, respectively).

[^3]:    ${ }^{5}$ This conclusion is reached by using the $F$ goodness of fit test for $k$-equal expectations per day. We obtained $F=264.59, p=0.000$ for departure delays, and $F=372.14, p=0.000$ for arrival delays. In both cases, we reject the null hypothesis.
    ${ }^{6}$ This is due to the fact that in the month analysed (December, 2014) the traffic was more important on the 23rd (Tuesday). To analyse this issue, we then tested the differences in proportions on departure delays in two ways. First, one for the 23 rd versus all other days ( $z=-46.62, p=0.000$ ). And second, one for the 23rd versus all other Tuesdays in the month $(z=-33.81, p=0.000)$. Both results indicate that there was a significant difference in delays between Tuesday 23rd and the other weekdays during December. In the same way, we studied the arrival delays and the results for the two tests were $z=-52.30, p=0.000$ and $z=-36.27$, $p=0.000$, respectively, and thus the same conclusions were reached.

[^4]:    *** indicates $1 \%$ significance or relevance level.
    ${ }_{*}^{* *}$ indicates $5 \%$ significance or relevance level.

    * indicates $10 \%$ significance or relevance level.

